Confidence Intervals II

18.05 Spring 2014

Jeremy Orloff and Jonathan Bloom

Agenda

- Polling: estimating p in Bernoulli(p).
- CLT $\Rightarrow$ large sample confidence intervals for the mean.
- Three views of confidence intervals.
- Constructing a confidence interval without normality: the exact binomial confidence interval for p

Polling: a binomial proportion confidence interval

Data $x_1, \ldots, x_n$ from a Bernoulli(p) distribution with p unknown.

A 'conservative normal'† $(1-\alpha)$ confidence interval for p is given by

$$\left[\bar{x} - \frac{z_{\alpha/2}}{2\sqrt{n}},\ \bar{x} + \frac{z_{\alpha/2}}{2\sqrt{n}} \right].$$

Proof uses the CLT and the observation $\sigma = \sqrt{p(1-p)} \leq 1/2$.

Political polls often give a margin-of-error of $\pm 1/\sqrt{n}$. This **rule-of-thumb** corresponds to a 95% confidence interval:

$$\left[\bar{x} - \frac{1}{\sqrt{n}},\ \bar{x} + \frac{1}{\sqrt{n}} \right].$$

†There are many types of binomial proportion confidence intervals.

http://en.wikipedia.org/wiki/Binomial_proportion_confidence_interval

Concept question: overnight polling

During the presidential election season, pollsters often do 'overnight polls' and report a 'margin of error' of about $\pm 5\%$.

The number of people polled is in which of the following ranges?

(a) 0 – 50
(b) 50 – 100
(c) 100 – 300
(d) 300 – 600
(e) 600 – 1000

National Council on Public Polls: Press Release, Sept 1992

"The National Council on Public Polls expressed concern today about the current spate of overnight Presidential polls. [...] Overnight polls do a disservice to both the media and the research industry because of the considerable potential for the results to be misleading. The overnight interviewing period may well mean some methodological compromises, the most serious of which is.."

...what?

National Council on Public Polls: Press Release, Sept 1992

"The National Council on Public Polls expressed concern today about the current spate of overnight Presidential polls. [...] Overnight polls do a disservice to both the media and the research industry because of the considerable potential for the results to be misleading. The overnight interviewing period may well mean some methodological compromises, the most serious of which is.."

...what?

"...the inability to make callbacks, resulting in samples that do not adequately represent such groups as single member households, younger people, and others who are apt to be out on any given night. As overnight polls often result in findings that are less reliable than those from more carefully conducted polls, if the media reports them, it should be with great caution."

http://www.ncpp.org/?q=node/42

Board question

A $(1-\alpha)$ confidence interval for p is given by

$$\left[\bar{x} - \frac{z_{\alpha/2}}{2\sqrt{n}},\ \bar{x} + \frac{z_{\alpha/2}}{2\sqrt{n}}\right].$$

1. How many people would you have to poll to have a margin of error of .01 with 95% confidence? (You can do this in your head.)
2. How many people would you have to poll to have a margin of error of .01 with 80% confidence. (You'll want R or a table here.)
3. If $n = 900$, compute the 95% and 80% confidence intervals for p.

Non-normal data

Suppose the data $x_1, x_2, \ldots, x_n$ is drawn from a distribution $f(x)$ that may not be normal or even parametric, but has finite mean, variance.

A version of the CLT says that for large n, the sampling distribution of the studentized mean is approximately standard normal:

$$\frac{\bar{x} - \mu}{s/\sqrt{n}} \quad \approx \quad \mathrm{N}(0, 1)$$

So for large n the $(1 - \alpha)$ confidence interval for μ is approximately

$$\left[\bar{x} - \frac{s}{\sqrt{n}} \cdot z_{\alpha/2}, \;\; \bar{x} + \frac{s}{\sqrt{n}} \cdot z_{\alpha/2}\right]$$

where $z_{\alpha/2}$ is the $\alpha/2$ critical value for $\mathrm{N}(0, 1)$.

This is called the *large sample confidence interval*.

Review: confidence intervals for normal data

Suppose the data $x_1, \ldots, x_n$ is drawn from $\mathrm{N}(\mu, \sigma^2)$

Confidence level $= 1 - \alpha$

- z confidence interval for the mean (σ known)

$$\left[\overline{x} - \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}}, \quad \overline{x} + \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}}\right]$$

- t confidence interval for the mean (σ unknown)

$$\left[\overline{x} - \frac{t_{\alpha/2} \cdot s}{\sqrt{n}}, \quad \overline{x} + \frac{t_{\alpha/2} \cdot s}{\sqrt{n}}\right]$$

- χ^2 confidence interval for σ^2

$$\left[\frac{n-1}{c_{\alpha/2}} s^2, \quad \frac{n-1}{c_{1-\alpha/2}} s^2\right]$$

- t and χ^2 have $n - 1$ degrees of freedom.

Three views of confidence intervals

View 1: Define/construct CI using a standardized point statistic.

View 2: Define/construct CI based on hypothesis tests.

View 3: Define CI as any interval statistic satisfying a formal mathematical property.

View 1: using a standardized point statistic

Example. $x_1 \ldots, x_n \sim \mathrm{N}(\mu, \sigma^2)$, where σ is known.

The *standardized* sample mean follows a standard normal distribution.

$$z \;=\; \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} \;\sim\; \mathrm{N}(0,1)$$

Therefore:

$$P(-z_{\alpha/2} < \frac{\overline{x} - \mu}{\sigma/\sqrt{n}} < z_{\alpha/2} \mid \boldsymbol{\mu}) \;=\; 1 - \alpha$$

Unwind to:

$$P(\overline{x} - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \;<\; \mu \;<\; \overline{x} + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \mid \boldsymbol{\mu}) \;=\; 1 - \alpha$$

This is the $(1 - \alpha)$ confidence interval:

$$\overline{x} \;\pm\; z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

Think of it as $\overline{x} \pm$ error

View 1: other standardized statistics

The t and χ^2 statistics fit this paradigm as well:

$$t = \frac{\overline{x} - \mu}{s/\sqrt{n}} \sim t(n-1)$$

$$X^2 = \frac{(n-1)s^2}{\sigma^2} \sim \chi^2(n-1)$$

View 2: using hypothesis tests

Set up: Unknown parameter θ. Test statistic x.

For any value θ_0, we can run an NSHT with null hypothesis

$$H_0 : \theta = \theta_0$$

at significance level α.

Definition. Given x, the $(1 - \alpha)$ confidence interval contains all θ_0 which are not rejected when they are the null hypothesis.

Definition. A type 1 CI error occurs when the confidence interval does not contain the true value of θ.

For a $1 - \alpha$ confidence interval, the type 1 CI error rate is α.

Board question: exact binomial confidence interval

Definition. Given x, the $(1-\alpha)$ confidence interval contains all θ_0 which are not rejected when they are the null hypothesis.

Use view 2 and this table of binomial(8,θ) probabilities to:

1. find the rejection region with significance level 0.10 for each value of θ –put values of x with the smallest probability into the rejection region.
2. Given $x = 7$, find the 90% confidence interval for θ.
3. Repeat for $x = 4$.

θ/x	0	1	2	3	4	5	6	7	8
.1	0.430	0.383	0.149	0.033	0.005	0.000	0.000	0.000	0.000
.3	0.058	0.198	0.296	0.254	0.136	0.047	0.010	0.001	0.000
.5	0.004	0.031	0.109	0.219	0.273	0.219	0.109	0.031	0.004
.7	0.000	0.001	0.010	0.047	0.136	0.254	0.296	0.198	0.058
.9	0.000	0.000	0.000	0.000	0.005	0.033	0.149	0.383	0.430

Solution

For each θ, the non-rejection region is blue, the rejection region is red.
In each row, the rejection region has probability at most $\alpha = .10$.

θ/x	0	1	2	3	4	5	6	7	8
.1	0.430	0.383	0.149	0.033	0.005	0.000	0.000	0.000	0.000
.3	0.058	0.198	0.296	0.254	0.136	0.047	0.010	0.001	0.000
.5	0.004	0.031	0.109	0.219	0.273	0.219	0.109	0.031	0.004
.7	0.000	0.001	0.010	0.047	0.136	0.254	0.296	0.198	0.058
.9	0.000	0.000	0.000	0.000	0.005	0.033	0.149	0.383	0.430

For $x = 7$ the 90% confidence interval for p is $[.7, .9]$.
These are the blue entries in the $x = 7$ column.

For $x = 4$ the 90% confidence interval for p is $[.3, .7]$.

View 3: formal

Recall: An interval statistic is an interval I_x computed from data x.

This is a random interval because x is random.

Suppose x is drawn from $f(x|\theta)$ with unknown parameter θ.

Definition:
A $(1-\alpha)$ confidence interval for θ is an interval statistic I_x such that

$$P(I_x \text{ contains } \theta \mid \theta) \;=\; 1-\alpha$$

for all possible values of θ (and hence for the true value of θ).

Note: equality in this equation is often relaxed to $\geq$ or $\approx$.
$=$: z, t, χ^2
$\geq$: rule-of-thumb and exact binomial (polling)
$\approx$: large sample confidence interval

18.05 Introduction to Probability and Statistics

Spring 2014